The Plant Journal (2003) **34**, 605–621

Sequence-based alignment of sorghum chromosome 3 and rice chromosome 1 reveals extensive conservation of gene order and one major chromosomal rearrangement

Patricia E. Klein[1,2,*], Robert R. Klein[3], Julia Vrebalov[4] and John E. Mullet[1,5]
[1]*Institute for Plant Genomics and Biotechnology, Texas A&M University, College Station, TX 77843, USA,*
[2]*Department of Horticultural Sciences, Texas A&M University, College Station, TX 77843, USA,*
[3]*USDA-ARS, Southern Plains Agricultural Research Center, College Station, TX 77845, USA,*
[4]*Boyce Thompson Institute, Ithaca, NY 14853, USA, and*
[5]*Department of Biochemistry and Biophysics, Texas A&M University, College Station, TX 77843, USA*

Received 31 December 2002; revised 7 February 2003; accepted 18 February 2003.
[*]For correspondence (fax +1 979 862 4790; e-mail pklein@tamu.edu).

Summary

The completed rice genome sequence will accelerate progress on the identification and functional classification of biologically important genes and serve as an invaluable resource for the comparative analysis of grass genomes. In this study, methods were developed for sequence-based alignment of sorghum and rice chromosomes and for refining the sorghum genetic/physical map based on the rice genome sequence. A framework of 135 BAC contigs spanning approximately 33 Mbp was anchored to sorghum chromosome 3. A limited number of sequences were collected from 118 of the BACs and subjected to BLASTX analysis to identify putative genes and BLASTN analysis to identify sequence matches to the rice genome. Extensive conservation of gene content and order between sorghum chromosome 3 and the homeologous rice chromosome 1 was observed. One large-scale rearrangement was detected involving the inversion of an approximately 59 cM block of the short arm of sorghum chromosome 3. Several small-scale changes in gene collinearity were detected, indicating that single genes and/or small clusters of genes have moved since the divergence of sorghum and rice. Additionally, the alignment of the sorghum physical map to the rice genome sequence allowed sequence-assisted assembly of an approximately 1.6 Mbp sorghum BAC contig. This streamlined approach to high-resolution genome alignment and map building will yield important information about the relationships between rice and sorghum genes and genomic segments and ultimately enhance our understanding of cereal genome structure and evolution.

Keywords: comparative mapping, sorghum, rice genomics.

Introduction

Over the past 5 years, the genome sequences of numerous key eukaryotic organisms have been completed (Adams *et al.*, 2000; The Arabidopsis Genome Initiative, 2000; The *C. elegans* Sequencing Consortium, 1998; Venter *et al.*, 2001). However, even though the cost of genome sequencing continues to decline, most eukaryotic genomes, especially those of plants and animals of agricultural importance, will not be completely sequenced in the near future. Fortunately, the genetic make-up and genome organization of related species is often sufficiently conserved to allow the alignment of genomes of species not targeted for sequencing with completely sequenced reference genomes (Bennetzen, 2000). Genome alignment enables research communities working on non-target species to predict the presence of genes in trait loci, locate expressed sequence tags (ESTs) on aligned maps, build physical maps at lower cost, and conduct focused comparative genome analyses with greater confidence. While the value of alignment is clear, criteria for identifying optimal pairs of non-target and reference genomes for alignment need to be clarified and methods for aligning genomes improved.

Sorghum is the fifth most important cereal worldwide (Doggett, 1988). Sorghum has unusual tolerance to adverse environments and is a potential source of many beneficial

genes for agriculture (Doggett, 1988). Currently, the sorghum genome is not targeted for complete sequencing even though the genome is relatively small (740 Mbp), and a high-density genetic map linked to an emerging BAC-based physical map is available (Klein *et al.*, 2000; Menz *et al.*, 2002). Sorghum is a member of the grass family, Poaceae, which includes other staple cereal crops, such as rice, maize, barley, oat, rye, millets, and wheat. Members of this family provide approximately 60% of the world's food production and thus are of great agronomic and economic importance.

The grass genomes exhibit an enormous variability in DNA content, ranging from approximately 430 Mbp in diploid rice to approximately 16 000 Mbp in hexaploid wheat (Arumuganathan and Earle, 1991). This variation in genome size is largely caused by differences in ploidy and the amount of repetitive DNA, primarily retrotransposons, present in centromeric regions and inserted between islands of gene clusters in some genomes (Panstruga *et al.*, 1998; SanMiguel and Bennetzen, 1999). In spite of the enormous differences in genome size, comparative genetic mapping using common DNA markers has revealed that the relative location of markers and mapped genes show remarkable conservation among the cereals. These results suggest that there are large stretches of gene collinearity within grass genomes, although numerous segmental duplications and deletions have been observed (Binelli *et al.*, 1992; Chen *et al.*, 1997; Helentjaris *et al.*, 1988; Hulbert *et al.*, 1990; Peng *et al.*, 1999; Ventelon *et al.*, 2001). Although comparative mapping studies have revealed conservation of marker order along large chromosomal segments, these results do not necessarily imply that all the genes located between these markers are collinear (Paterson *et al.*, 1996). Exceptions to collinearity, observable only at the DNA sequence level, have been reported, and in many cases, appear to be the result of gene amplification, gene movement and inversion, as well as retrotransposition (for review, see Bennetzen, 2000; Dubcovsky *et al.*, 2001; Keller and Feuillet, 2000; Song *et al.*, 2002).

Most present-day comparative maps of grass species are of limited resolution as they are based on a finite number of cross-hybridizing DNA probes that must be polymorphic in the species being analyzed. In contrast, high-resolution gene sequence-based comparative maps have a resolution that far exceeds the current RFLP probe-based comparative maps. Therefore, the aim of this study was to explore methods for high-resolution alignment of the sorghum and rice genomes. To this end, we conducted a sequence scan of 118 BACs mapped across sorghum chromosome 3 to facilitate alignment to the rice genome sequence and developed a method for targeted contig building based on the resulting sorghum–rice genome alignment.

Results

Integrated genetic and physical mapping of sorghum chromosome 3

In a previous study, we detailed the methodology being utilized for the construction of an integrated genetic and physical map of the sorghum genome (Klein *et al.*, 2000). This included BAC DNA fingerprinting for DNA contig construction and six-dimensional BAC DNA pooling combined with AFLP technology for the integration of genetic markers into BAC contigs. In the present study, we utilized this methodology to construct a framework integrated genetic and physical map of sorghum chromosome 3 (previously designated linkage group C, Menz *et al.*, 2002). One hundred and ninety-two AFLP primer combinations (+3/+3) previously used in the construction of the high-density sorghum genetic map were used to screen BTx623 BAC DNA pools (Klein *et al.*, 2000). Additionally, the BAC DNA pools were screened with primers for five STSs corresponding to sorghum RFLPs and 10 sorghum SSRs. From this analysis, 731 BACs were anchored to 135 loci located every 2–5 cM (approximately) across chromosome 3. On average, each marker identified 2.6 BAC clones in agreement with our previous results (Klein *et al.*, 2000). From previous fingerprint analysis, 711 of these BACs were assembled into 120 contigs while 20 remained as singletons (Klein *et al.*, 2000). Each contig contained an average of 5.9 BAC clones. To confirm the accuracy of BAC identification using marker analysis of the BAC DNA pools, 47 individual BAC clones identified from the DNA pools as positive for a given marker were analyzed for the presence of that marker. Forty-six of the 47 individual clones amplified a band corresponding to the appropriate genetic marker, indicating a false-positive rate of less than 3% (data not shown).

Criteria for sequence-based alignment of the sorghum and rice genome maps

Utilizing the framework integrated map of sorghum chromosome 3 described above, we set out to test the utility of sequence-based alignment between sorghum chromosome 3 and the rice genome sequence. Five BACs that mapped at different locations in the distal regions of sorghum chromosome 3 were chosen for an initial targeted analysis. Ninety-six *Eco*RI/*Xho*I subclones from each BAC were selected and sequenced from the *Eco*RI cloning site. Following the removal of sequences containing only cloning vector or those with low-quality scores, approximately 83 sorghum sequences with an average read length of 779 bp were obtained from each BAC. PHRAP analysis of these sequences showed that 10–33 unique sequences were obtained per BAC clone. The results of BLASTX and

Table 1 Sequence scanning of sorghum BACs and alignment to rice chromosome 1

Sorghum chr. 3 location[a]	Genetic marker	Sb BAC	No. of unique seq.[b]	Protein accession no.[c]	Protein designation	BLASTX expected value[d]	Rice BAC/PAC accession no.[c]	BLASTN expected value[e]	Rice location[g]
28.6–31.6	*Xtxa2598*	122c5	33	BAB64774-(2)	Unknown	3×10^{-42} [f]	AP003282-(2)	8×10^{-22} [f]	1–16.4
				BAB64770-(1)	Hypothetical	7×10^{-13}	AP003282-(2)	2×10^{-18}	1–16.4
				AAM69848-(1)	Putative alliin lyase	6×10^{-29}	AP003339-(1)	6×10^{-67}	1–16.4
				BAB92519-(1)	Putative alliinase	9×10^{-41}	AP003339-(1)	2×10^{-30}	1–16.4
				BAB21145-(1)	Hypothetical	2×10^{-13}	AP002899-(1)	4×10^{-28}	1–126.2
				BAB91929-(2)	Putative receptor-like kinase	1×10^{-71}	AP003768-(2)	1×10^{-68}	1–127.3
				BAB91928-(1)	Putative dehydrogenase	1×10^{-103}	AP003768-(1)	0.0	1–127.3
				BAB91943-(2)	Putative kinesin-like	7×10^{-33}	AP003768-(2)	6×10^{-30}	1–127.3
158.6–161.7	*Xtxp38*	7b9	10	P94044-(1)	Ferredoxin VI	2×10^{-33}	AP003794-(1)	7×10^{-85}	1–145.3
				BAB91757-(1)	Putative 60S ribosomal	7×10^{-11}	AP003794-(1)	3×10^{-48}	1–145.3
				BAB90565-(1)	Putative chitinase	2×10^{-70}	AP003794-(1)	5×10^{-92}	1–145.3
169.2–173.4	*Xtxa2667*	112d8	21	BAB92564-(2)	Putative zinc finger	7×10^{-59}	AP003380-(2)	3×10^{-78}	1–157.6
				BAB92572-(2)	Unknown	2×10^{-13}	AP003380-(2)	1×10^{-40}	1–157.6
				BAB92573-(2)	Mg2 transporter-like	4×10^{-26}	AP003380-(2)	2×10^{-23}	1–157.6
				AAL61542-(1)	Isoflavone reductase-like	1×10^{-11}	AP003545-(1)	7×10^{-36}	6–65.8
173.4–175.1	*Xtxa3719*	181g10	13	BAB86148-(1)	Unknown	6×10^{-19}	AP003451-(1)	1×10^{-105}	1–158.5
				AAM61500-(1)	Hydrolase-like	5×10^{-13}	AP003437-(1)	1×10^{-12}	1–159.0
				BAB86097-(1)	Unknown	3×10^{-10}	AP003437-(1)	6×10^{-6}	1–159.0
				BAB86107-(2)	Sucrose-phosphate synthase	2×10^{-40}	AP003437-(1)	1×10^{-43}	1–159.0
				BAB86109-(1)	Unknown	4×10^{-21}	AP003437-(1)	5×10^{-21}	1–159.0
175.1	*Xtxa3891*	250h7	20	BAB86123-(1)	Hypothetical	2×10^{-46}	AP003437-(1)	1×10^{-11}	1–159.0
				BAB86119-(1)	Hypothetical	5×10^{-40}	AP003437-(1)	1×10^{-59}	1–159.0
				BAB86175-(2)	Putative alpha-glucosidase	2×10^{-21}	AP003707-(2)	2×10^{-39}	1–159.6
				BAB86174-(1)	Hypothetical	3×10^{-61}	AP003707-(1)	1×10^{-37}	1–159.6

[a]Position, in centimorgans, on sorghum chromosome 3.
[b]Unique sequences remaining after removal of those containing vector only and low quality scores.
[c]Numbers in paranthesis next to accession numbers indicate the number of sorghum sequences with homology to that gene or rice BAC/PAC.
[d]BLASTX to the nr database.
[e]BLASTN to the nt or htg databases.
[f]When a gene or BAC/PAC share homology to more than one sorghum sequence, the highest expect value obtained is shown.
[g]Rice chromosome number and position in centimorgans.

BLASTN analyses for these sequences are shown in Table 1. Three criteria were utilized to assess the value of sorghum and rice sequence matches, and all three criteria had to be satisfied before sorghum sequences were used to align the chromosomes. These criteria were: (i) the sorghum sequence had to show sequence similarity to a putative, hypothetical or known protein coding sequence in the non-redundant database at an expected value greater than 10^{-5}; (ii) the sorghum sequence had to match a rice BAC sequence in the non-redundant or high-throughput genomic sequence databases at an expected value greater than 10^{-5}; and (iii) there had to be a significant overlap between the sequences satisfying the first two criteria. In addition, each sequence that could potentially align sorghum and rice chromosomes was examined for homology to large gene families, especially those associated with known transposable elements. Those sequences associated with transposable elements were eliminated from consideration, whereas sequences from large protein families were examined carefully to see if they would allow unique alignments to be established.

Approximately 25–53% of the sequences from each sorghum BAC showed significant similarity with protein coding sequences (BLASTX, *E*-value $> 10^{-11}$) as well as rice BAC sequences (BLASTN, *E*-value $> 6 \times 10^{-6}$, Table 1). Of the 97 independent sequences collected from the five sorghum BACs, 32 showed similarity with 24 different genes (Table 1). Twenty-one of the genes with homology to sorghum sequences were from rice, and the remaining three genes were from maize, wheat, or *Arabidopsis*. The 21 rice genes having similarity with sorghum BAC sequences were located on 10 different rice BACs (Table 1). Examination of the integrated genetic and physical map of rice (http://rgp.dna.affrc.go.jp/cgi-bin/statusdb/stattable.pl?chr = 1&lab = = RGP) revealed that nine of the 10 rice BACs were located on rice chromosome 1 and seven of these BACs were collinear with the five sorghum BACs mapped to chromosome 3 (Table 1).

Although most of the genic sequences from the five sorghum BACs were collinear between sorghum chromosome 3 and rice chromosome 1, some exceptions to collinearity were observed (Table 1, denoted in red). For example, sorghum BAC 122c5 contained sequence similarity with genes, and rice BACs mapped to two different positions within rice chromosome 1. These rice BACs mapped to a position at 16.4 cM on rice chromosome 1, which was collinear with the mapped position of BAC 122c5 at approximately 28.6–31.6 cM on sorghum chromosome 3 as well as with a non-collinear position between approximately 126.2 and 127.3 cM. To ensure that sorghum BAC 122c5 was not contaminated with a second BAC from sorghum chromosome 3, six independent colonies from this stock were isolated and analyzed by AFLP analysis. All six clones showed identical AFLP banding patterns with two different +3/+3 primer combinations, indicating that the stock of BAC 122c5 represented a single clone (data not shown). Finally, one sequence from sorghum BAC 112d8 shared similarity with a rice BAC, which mapped to a position on rice chromosome 6, suggesting possible translocation of the isoflavone reductase-like gene in sorghum relative to rice (Table 1).

The results of this preliminary examination revealed that, on average, 33% of the independent sorghum sequences (32/97) had significant similarity with genes present in the non-redundant database and also with rice BACs. Based on these results, we reasoned that sample sequencing 16–32 *Eco*RI/*Xho*I subclones per BAC would provide sufficient sequence information to permit alignment of BACs mapped in sorghum to the rice genome.

Sequence-based alignment of sorghum chromosome 3 to rice chromosome 1

One hundred and eighteen BACs anchored to 109 loci distributed across sorghum chromosome 3 were selected for sequence scanning and alignment to rice chromosome 1. Following PHRAP analysis, the number of unique sequences obtained per BAC ranged from 3 to 18 with an average of 9 at a read length of 825 bp. The sequences were compared to the non-redundant (BLASTX and BLASTN) and high-throughput genomic sequence (htgs, BLASTN) databases, and alignment to the rice genome was determined following the criteria described above.

One hundred and sixty different sequences (GenBank accession nos. BZ412839–BZ412998) from 63 BACs mapping to 54 loci on sorghum chromosome 3 showed sequence similarity with 67 BACs from rice chromosome 1 (Table 2). Based on an analysis of 145 independent sequences, 59 of the 63 sorghum BACs were collinear with the minimum tiling path of BACs from rice chromosome 1, providing strong evidence for a syntenic relationship between sorghum chromosome 3 and rice chromosome 1. The remaining four sorghum BACs exhibited sequence similarity with BACs from rice chromosome 1, although the position of these sorghum BACs was not collinear with the other rice BACs from this chromosome (Table 2, BACs 213f6, 92f9, 159b4, and 68g5, denoted in blue; accession nos. BZ412993–BZ412998). A limited number of unique sequences from 11 of the 59 sorghum BACs that exhibited collinearity with rice chromosome 1 also contained sequences that aligned best with rice BACs located in another part of the rice genome (Table 2, sequences denoted in red; accession nos. BZ412855–BZ412859, BZ412861, BZ412931, BZ412965, BZ412975, BZ412999–BZ413010).

Five BACs from sorghum chromosome 3 did not contain sequences having similarity with rice chromosome 1 but instead exhibited sequence similarity with BACs from other rice chromosomes (data not shown). AFLP analysis of these five BACs confirmed the presence of the respective AFLP marker within each BAC, suggesting that these BACs had been correctly mapped to sorghum chromosome 3 (data not shown). Thus, based on the scanning of BAC sequences, these five sorghum BACs were best aligned with regions of the rice genome other than rice chromosome 1.

Fifty of the 118 sorghum BACs examined did not show sequence similarity with any rice genic sequence meeting our criteria for alignment. Nearly one-half of these (24/50) were anchored to markers located within the heterochromatic region of sorghum chromosome 3 (N. Islam-Faridi, personal communication), and the majority of the sequences obtained from these BACs had similarity to repetitive elements (i.e. gag-pol polyprotein, retrotransposon-like elements, and reverse transcriptase). Heterochromatic regions that are largely devoid of genes reduce the efficiency of alignment based on our criteria for gene sequence matches between sorghum and rice.

Figure 1 depicts the alignment of BACs mapped on sorghum chromosome 3 and rice chromosome 1 based on the results of sequence scanning. In Figure 1, sorghum chromosome 3 has been re-drawn relative to our initial mapping study (Menz *et al.*, 2002) because recent cytologic characterization of this chromosome indicates that the centromere is located between positions approximately 76–80 cM (N. Islam-Faridi, personal communication). Therefore, the order of markers on this chromosome has been inverted such that the short arm of the chromosome is above the centromere and the long arm below. This re-ordering of markers along sorghum chromosome 3 relative to the centromere is consistent with the integrated cytologic and genetic map of rice chromosome 1 (Cheng *et al.*, 2001). Examination of Figure 1 reveals that the alignment of mapped BACs between sorghum chromosome 3 and rice chromosome 1 is largely collinear at this level of resolution with the exception of one major chromosomal

Table 2 Alignment of syntenic BACs between sorghum chromosome 3 and rice chromosome 1

Sorghum chr. 3 location[a]	Genetic marker	Sb BAC	Rice BAC accession no.[b]	BLASTN expected value[c]	Protein accession no.[d]	BLASTX expected value[e]	Rice location[f]
0.0	*Xtxa3395*	154b9	AP002900-(1)	1×10^{-160}	BAB32724	3×10^{-90}	1–32.4–36.9
9.0–11.9	*Xtxa2079*	228d5	AP003052-(1)	5×10^{-43}	BAB92155	2×10^{-46}	1–28.9–29.7
16.7	*Xtxa376*	163a2	AP002836-(1)	2×10^{-72}	BAB03418	1×10^{-29}	1–25.4
28.6–31.6	*Xtxa2598*	122c5	AP003282-(3)	8×10^{-22}	BAB64774	2×10^{-11}	1–16.4
			AP003339-(2)	6×10^{-67}	AAM69848	6×10^{-29}	1–16.4
			AP002899-(1)	4×10^{-28}	BAB21145	2×10^{-13}	1–126.2
			AP003768-(5)	0.0	BAB91928	1×10^{-103}	1–127.3
31.6	*Xtxa2614*	230a10	AP003301-(1)	1×10^{-28}	BAB39900	4×10^{-33}	1–16.1
			AP003339-(1)	2×10^{-20}	BAB39913	9×10^{-13}	1–16.1
33.3	*Xtxa6004*	212a2	AP002539-(2)	2×10^{-82}	BAB08208	4×10^{-58}	1–13.1
33.3–37.3	*Xtxa2328*	74a7	AP003209-(1)	2×10^{-17}	BAB64789	4×10^{-13}	1–13.1–16.1
			AP004145-(1)	5×10^{-43}	BAB86275	1×10^{-73}	1–114.1–116.5
33.3–37.3	*Xtxa2328*	86a2	AP003209-(6)	2×10^{-39}	BAB64789	6×10^{-59}	1–13.1–16.1
			AP001551-(1)	3×10^{-13}	BAA92961	8×10^{-16}	1–28.9
			AL731613-(1)	2×10^{-33}	NP_187305	6×10^{-9}	4-n/a
37.3	*Xtxa3003*	167d8	AP003045-(1)	3×10^{-19}	BAB03621	2×10^{-25}	1–12.2
			AP002540-(1)	7×10^{-17}	BAB03621	2×10^{-61}	1–12.2
37.3	*Xtxa3003*	213f6	AP003292-(1)	1×10^{-13}	BAB84438	4×10^{-28}	1–124.8
43.0–47.2	*Xtxa6085*	215g4	AP002860-(1)	7×10^{-33}	BAB18289	2×10^{-26}	1–8.7
43.0–47.2	*Xtxp215*	133h9	AP003214-(1)	5×10^{-15}	BAB64612	5×10^{-36}	1–10.9
			AP003214-(1)	9×10^{-17}	AAL14403	5×10^{-5}	1–10.9
			AP004632-(1)	9×10^{-8}	NP_199704	9×10^{-11}	8–114.4
			AP005582-(1)	3×10^{-32}	NP_564794	4×10^{-17}	9–0.8
53.7–55.7	*Xtxa2822*	162e5	AP002487-(1)	1×10^{-34}	BAB07943	3×10^{-16}	1–7.0
58.5	*Xtxa2260*	205c5	AP002969-(3)	0.0	BAB62641	4×10^{-74}	1–0.0
62.0	*Xtxa3837*	203a3	AP001278-(1)	8×10^{-14}	BAA92196	2×10^{-20}	1–43.2
			AP003769-(1)	1×10^{-6}	AAD38297	3×10^{-10}	6–122.8
			AC103550-(1)	2×10^{-36}	NP_177372	1×10^{-35}	3–135.1
65.5–68.7	*Xtxa2746*	92f9	AP003286-(2)	1×10^{-36}	BAB89808	1×10^{-11}	1–146.4
			AL607101-(1)	1×10^{-176}	NP_174466	1×10^{-102}	3-n/a
76.5–79.9	*Xtxa2646*	84f8	AP003722-(1)	1×10^{-12}	BAB92783	4×10^{-9}	1–62.5–63.9
76.5–79.9	*Xtxa3618*	177f5	AP003202-(1)	4×10^{-34}	BAB92229	2×10^{-21}	1–73.4–73.7
76.5–79.9	*Xtxa3688*	184b5	AP003372-(1)	1×10^{-31}	BAB89081	1×10^{-15}	1–86.0
76.5–82.3	*Xtxa3823*	159b4	AP003209-(1)	2×10^{-17}	BAB64789	4×10^{-13}	1–13.1–16.1
84.7–88.6	*Xtxa2768*	73c9	AP003546-(1)	9×10^{-23}	BAB32948	8×10^{-23}	1–93.1
84.7–88.6	*Xtxp205*	99h2	AP003048-(8)	1×10^{-65}	BAB55821	2×10^{-22}	1–95.7
91.0–94.4	*Xtxp31*	240d9	AP002844-(1)	7×10^{-36}	BAB21276	2×10^{-23}	1–100.4
105.2	*Xtxa4106*	114e8	AP002839-(2)	1×10^{-24}	BAB19086	2×10^{-30}	1–103.7
108.6–111.3	*Xtxa6219*	158h7	AP003855-(2)	4×10^{-28}	NP_197090	2×10^{-9}	1–106.2–107.6
112.0	*Xtxa2595*	103h7	AP003855-(2)	5×10^{-49}	BAB90468	1×10^{-46}	1–106.2–107.6
118.8–125.6	*Xtxp59*	86c2	AP004364-(1)	3×10^{-16}	BAB91748	1×10^{-26}	1–113.3
			AP003284-(2)	5×10^{-36}	BAB91748	4×10^{-26}	1–114.1
125.6–127.9	*Xtxa3940*	230g9	AP003760-(2)	1×10^{-18}	BAB90507	3×10^{-26}	1–114.1–116.5
125.6–127.9	*Xtxa2233*	204g12	AP003264-(1)	1×10^{-40}	BAB93325	4×10^{-83}	1–116.9
130.5–133.7	*Xtxa4131*	117d8	AP004363-(2)	7×10^{-64}	BAA35120	2×10^{-17}	1–118.9
			AC091532-(1)	6×10^{-9}	NP_196234	4×10^{-6}	3–128.0
			AC093956-(1)	2×10^{-30}	CAC85920	2×10^{-35}	5–111.6
133.7	*Xtxa4019*	68g5	AP003045-(1)	3×10^{-6}	BAB03629	6×10^{-53}	1–12.2
			AP002540-(1)	3×10^{-24}	BAB03605	3×10^{-14}	1–12.2
133.7–135.1	*Xtxa594*	122h8	AP003373-(1)	1×10^{-131}	BAC10738	8×10^{-96}	1–122.1
136.0	*Xtxa2984*	209f3	AP003410-(2)	6×10^{-39}	BAB39420	5×10^{-55}	1–123.2
			AL731881-(1)	8×10^{-27}	AAM53297	2×10^{-8}	12–100.9
136.0–140.4	*Xtxa6120*	131f7	AP003292-(1)	3×10^{-47}	BAB84414	6×10^{-63}	1–124.8
			AP003229-(2)	4×10^{-31}	NP_194683	5×10^{-9}	1–124.8

Table 2 continued

Sorghum chr. 3 location[a]	Genetic marker	Sb BAC	Rice BAC accession no.[b]	BLASTN expected value[c]	Protein accession no.[d]	BLASTX expected value[e]	Rice location[f]
140.4–142.9	*Xtxa428*	48c11	AP003229-(2)	4×10^{-31}	NP_194683	5×10^{-9}	1–124.8
			AP002972-(1)	2×10^{-54}	BAB55510	9×10^{-35}	1–126.5
			AP003768-(9)	0.0	BAB91928	1×10^{-110}	1–127.3
			AP003376-(4)	7×10^{-39}	BAC05580	3×10^{-15}	1–127.3
			AP005493-(1)	2×10^{-27}	AAD27911	2×10^{-17}	8–54.3
140.4–142.9	*Xtxa428*	44f4	AP003376-(1)	3×10^{-19}	BAB91943	5×10^{-10}	1–127.3
			AP003768-(2)	2×10^{-39}	BAB91930	4×10^{-19}	1–127.3
140.4–142.9	*Xtxa2999*	252c9	AP003376-(1)	4×10^{-49}	BAC05596	3×10^{-26}	1–127.3
145.9–147.7	*Xtxa443*	237g2	AP003274-(2)	8×10^{-51}	BAB61219	2×10^{-35}	1–130.1–132.0
145.9–147.7	*Xtxa3409*	95a10	AP003368-(1)	5×10^{-6}	BAB92557	1×10^{-28}	1–136.1
149.5	*Xtxa4063*	75c11	AP003280-(1)	1×10^{-138}	BAB89767	1×10^{-101}	1–136.6
			AP003345-(1)	2×10^{-32}	BAB90113	1×10^{-72}	1–136.6
149.5–150.7	*Xtxa3789*	172g12	AP003293-(1)	3×10^{-35}	BAB86413	7×10^{-52}	1–136.9
149.5–150.7	*Xtxa6125*	216g7	AP003232-(1)	1×10^{-33}	BAB92273	8×10^{-15}	1–136.9
149.5–150.7	*Xtxa6125*	165h12	AP003232-(2)	4×10^{-25}	BAB92273	2×10^{-9}	1–136.9
			AP003073-(1)	3×10^{-35}	BAB44078	2×10^{-10}	1–137.2
156.6–158.6	*Xtxa6040*	49a9	AP003240-(1)	1×10^{-100}	BAB86465	2×10^{-59}	1–143.7
			AP003241-(4)	5×10^{-52}	BAB93227	1×10^{-45}	1–143.7
156.6–158.6	*Xtxa6040*	92e1	AP003251-(1)	7×10^{-8}	BAB89569	5×10^{-13}	1–143.7
158.6–161.7	*Xtxp38*	114c4	AP003794-(1)	3×10^{-10}	BAB63718	2×10^{-18}	1–145.3
158.6–161.7	*Xtxp38*	7b9	AP003794-(3)	5×10^{-92}	BAB90565	2×10^{-70}	1–145.3
161.7	*Xtxa 2027*	233d11	AP003270-(1)	2×10^{-17}	BAB89788	3×10^{-7}	1–146.4
			AP003344-(1)	3×10^{-22}	BAC07356	2×10^{-45}	1–134.7–135.8
169.2–173.4	*Xtxp34*	232a8	AP003407-(1)	4×10^{-25}	BAB90185	6×10^{-39}	1–151–154.6
			AP003316-(2)	1×10^{-21}	BAC06263	8×10^{-37}	1–154.6
169.2–173.4	*Xtxa2032*	111e10	AP003316-(1)	1×10^{-21}	BAC06263	1×10^{-6}	1–154.6
			AC074232-(1)	3×10^{-7}	AAM18996	3×10^{-10}	10–61.7
169.2–173.4	*Xtxa6043*	78c12	AP003791-(1)	3×10^{-11}	NP_190330	2×10^{-11}	1–157.1
169.2–173.4	*Xtxa6043*	120h5	AP003791-(1)	9×10^{-60}	BAB90538	6×10^{-26}	1–157.1
			AP003380-(1)	8×10^{-76}	BAB92564	2×10^{-63}	1–157.6
169.2–173.4	*Xtxa2667*	112d8	AP003380-(6)	3×10^{-78}	BAB92564	7×10^{-59}	1–157.6
			AP003545-(1)	7×10^{-36}	AAL61542	1×10^{-11}	6–65.8
173.4–175.1	*Xtxa3719*	181g10	AP003451-(1)	1×10^{-105}	BAB86148	6×10^{-19}	1–158.5
			AP003437-(5)	1×10^{-43}	BAB86107	2×10^{-40}	1–159.0
175.1	*Xtxa3891*	250h7	AP003437-(2)	1×10^{-59}	BAB86119	5×10^{-40}	1–159.0
			AP003707-(3)	2×10^{-39}	BAB86175	2×10^{-21}	1–159.6
177.5–180.7	*Xtxa3987*	231g1	AP004672-(1)	2×10^{-33}	BAB90826	7×10^{-21}	1–156.9–161.5
177.5–180.7	*Xtxa3987*	140g1	AP004672-(2)	1×10^{-33}	BAB90826	4×10^{-21}	1–156.9–161.5
181.5	*Xtxa3904*	165d4	AP004332-(3)	4×10^{-31}	BAB89661	1×10^{-23}	1–161.5
182.3	*Xtxa4073*	174d8	AP004330-(1)	3×10^{-38}	BAB90754	1×10^{-15}	1–161.8
182.3–187.7	*Xtxp69*	132f4	AP003246-(2)	2×10^{-30}	BAB64184	1×10^{-28}	1–163.5
196.5	*isu-52*	126a12	AP004326-(2)	7×10^{-82}	BAB92874	3×10^{-18}	1–170.4–176.3
			AP003238-(1)	3×10^{-7}	BAC07382	3×10^{-10}	1–170.4–176.3
196.5	*isu-52*	137h12	AP003238-(2)	2×10^{-27}	BAB89013	6×10^{-21}	1–170.4–176.3
off-end	*Xtxa4138*	211e12	AP003448-(4)	3×10^{-38}	BAB85329	5×10^{-27}	1–178.1
			AP003277-(2)	7×10^{-24}	BAB63671	8×10^{-8}	1–181.8

[a]Sorghum chromosome 3 position in centimorgans.
[b]Number in parenthesis next to rice accession no. indicates the number of sorghum sequences with homology to that given BAC/PAC.
[c]BLASTN to the nt or htg databases.
[d]When more than one sorghum sequence from a given BAC showed homology to the same rice BAC/PAC, only the gene with the highest homology for those sequences is shown.
[e]BLASTX to the nr database.
[f]Rice chromosome number and position in centimorgans; n/a, genetic position not mapped.

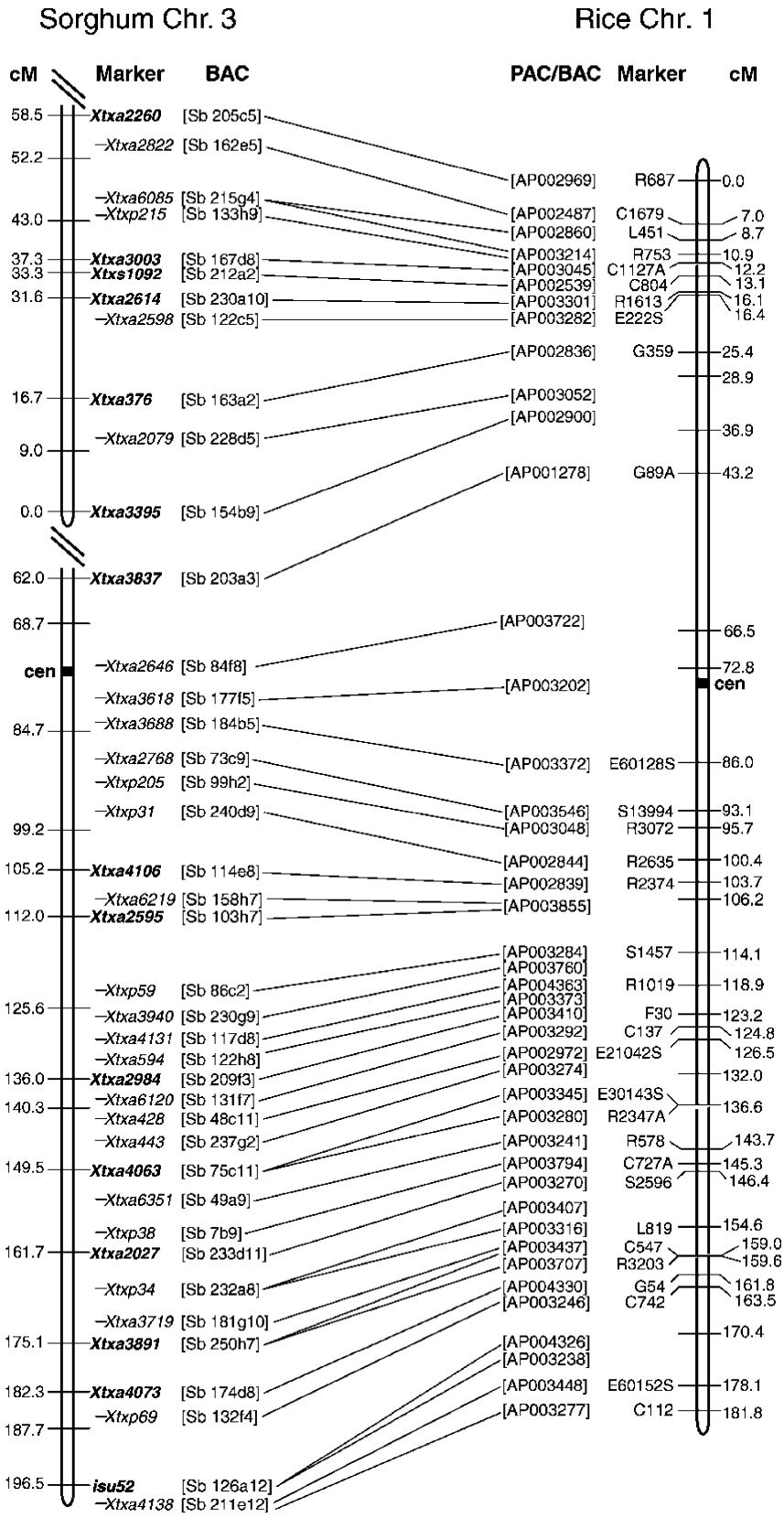

Figure 1. Sequence-based alignment of sorghum chromosome 3 and rice chromosome 1. The sorghum genetic map used is that of Menz *et al.* (2002), and the rice map is from the Rice Genome Research Program (RGP) (http://rgp.dna.affrc.go.jp/publicdata/geneticmap2000/chr01.html). The centromeres are denoted by black boxes in both maps. In the sorghum map, framework markers are in bold text and connected to the main bar adjacent to their map position. Markers 'placed' in bins between framework markers are in plain type and are not connected to the main bar. The sorghum BAC anchored to each loci and used for sample sequence analysis is listed in brackets next to its respective loci. Rice PACs/BACs (GenBank accession nos.) from the chromosome 1 physical map (http://rgp.dna.affrc.go.jp/cgi-bin/statusdb/stattable.pl?chr = 1&lab = RGP) are listed in brackets next to the rice genetic map. When the rice BAC/PAC contained a rice genetic marker, the locus is listed next to its map position along the rice map. Collinear sorghum and rice BACs are joined by horizontal lines. A break is shown in the short arm of sorghum chromosome 3 to depict the segmental inversion that occurred after the separation of sorghum and rice.

rearrangement. An inversion of a portion of sorghum chromosome 3 from positions 0–58.5 cM (markers *Xtxa3395*–*Xtxa2260*, respectively) has occurred in sorghum relative to rice (Figure 1). It is also evident that a large heterochromatic region of low gene density is located around the centromeres of these two chromosomes (Sasaki *et al.*, 2002; N. Islam-Faridi, personal communication). In the approximately 20 cM block surrounding the centromere of sorghum, only three gene sequence-based alignments between sorghum chromosome 3 and rice chromosome 1 were detected out of a total of 145 independent alignments in this study. In addition, a greater number of gene sequences were identified per sample sequence in BACs mapping closer to the ends of the chromosome compared to BACs nearer the centromere (Figure 1).

Using the rice genome sequence to accelerate physical map construction in sorghum

The establishment of sequence-based alignments between BACs mapped at intervals along sorghum chromosome 3 and the nearly complete sequence of rice chromosome 1 allowed the development of an efficient approach for filling gaps in the sorghum physical map. Previous fingerprinting of our BTx623-derived BAC libraries placed 48 of the 59 sorghum chromosome 3 BACs used for the sequence-based alignment into 41 different DNA contigs with 11 clones remaining in a pool of singletons (Klein *et al.*, 2000). The high stringency that was initially used for contig construction in FPC (10^{-14}) required that BAC clones share at least 60% overlap in their fingerprints before being automatically placed within a contig (Klein *et al.*, 2000). Furthermore, contigs were not merged with other contigs or singletons unless two independent analyses supported the overlap between clones (i.e. DNA fingerprinting and marker content mapping, Klein *et al.*, 2000). Sequence scans of BACs in this study provided an additional source of information for contig merging. For example, sequences obtained from the 59 BACs aligned to rice chromosome 1 provided evidence supporting at least 10 different merges between contigs and/or singletons. Direct evidence for four contig merges was obtained when sorghum BACs from two different contigs and/or singletons shared sequence similarity with the same rice gene located on the same rice chromosome 1 BAC. For example, ctg612 (anchored to marker *Xtxp34*) was manually merged with ctg3544 (anchored to marker *Xtxa2032*) after sequence scanning of BACs 232a8 (ctg612) and 111e10 (ctg3544) revealed that both BACs shared sequence similarity with a putative calreticulin gene from rice chromosome 1 BAC AP003316 (Table 2). In the other six cases, indirect evidence supporting contig merges was obtained when two BACs from two different contigs (or from the pool of singletons) shared sequence similarity with the same BAC from rice chromosome 1. However, as the sequence shared by the rice BAC and the two sorghum BACs did not overlap and was not from a genic region, this data only provided tentative support for merging. Contigs, whose merges were supported indirectly, were marked in our FPC database for follow-up analysis.

The sequence-based alignment of BACs mapped on sorghum chromosome 3 and the sequence of rice chromosome 1 allowed genes located in gaps present in the sorghum physical map to be identified based on the predicted collinearity of most genes in the target intervals. The identification of sorghum genes that are located in physical map gaps provides the necessary information to identify those BACs that encode these genes and hence reside in these gaps. The approach taken was to identify rice gene sequences predicted to lie within a gap in the physical map of sorghum and search the sorghum EST database for orthologous gene sequences. The resulting sorghum EST sequences were used to design gene-specific PCR primers that could be used to screen for sorghum BACs containing these sequences. The BACs identified using this method were fingerprinted, along with the previously mapped BACs flanking the gap in the physical map, to determine their order, degree of overlap, and potential for filling gaps in the sorghum physical map.

This approach was tested on a targeted region of sorghum chromosome 3 between positions approximately 169.2 and 180.7 cM. Initial sequence scanning of 12 sorghum BACs mapped to nine genetic loci in this interval allowed alignment of this region of sorghum chromosome 3 to rice chromosome 1 between positions approximately 151.0 and 161.5 cM. The minimum tiling path spanning this region of rice chromosome 1 consists of 14 BACs with two small gaps and spans approximately 1.6–1.7 Mbp (http://rgp.dna.affrc.go.jp/cgi-bin/statusdb/stattable.-pl?chr = 1&lab = RGP). The aligned region in sorghum contained 32 BACs organized into six contigs and one singleton. In the first step toward filling the six gaps in this region of the sorghum map, the sequence of each of the 14 rice BACs was analyzed using BLAST analysis to the sorghum EST database. The BLASTN and TBLASTX analyses allowed us to identify sorghum ESTs homologous to genes encoded by each rice BAC including genes that should reside in gaps within the sorghum physical map. Sorghum ESTs that appeared to be single or low copy from this analysis were further screened against the non-redundant database (BLASTN) to ensure that each sorghum EST exhibited strong sequence similarity only with the appropriate BAC(s) on rice chromosome 1. Using this screening approach, 29 low-copy sorghum ESTs with homologs on rice chromosome 1 BACs (positions approximately 151.0–161.5 cM) were selected and utilized for PCR-based STS screening of six-dimensional BAC DNA pools derived from the two sorghum genotypes, BTx623 and IS3620C.

Table 3 Sorghum ESTs with homologs in rice chromosome 1 BACs/PACs

Sorghum EST accession no.	Rice BAC/PAC accession no.[a]	Rice chr.1 location	Nucleotide position in rice[b]	BLASTN expected value	Collinear in sorghum chr. 3[c]	Sorghum BAC[d]
BG103709	AP003349	154.6	16813–17199 (+/+)	1×10^{-53}	n/a	
BE600064	AP003349	154.6	55472–56057 (+/+)	1×10^{-135}	Yes	111e10
BG488122	AP003349	154.6	80903–81370 (+/–)	1×10^{-78}	Yes	111e10
BG050777	AP003418	155.2	40682–40990 (+/–)	6×10^{-57}	Yes	77G16
BM324945	AP003418	155.2	65930–66454 (+/–)	3×10^{-40}	Yes	79 N7
BE598683	AP003418	155.2	98316–100106 (+/+)	4×10^{-27}	n/a	
BE358182	AP003418	155.2	141561–142677 (+/–)	2×10^{-32}	Yes	79 N7
AW746672	AP003313	157.1	86030–87678 (+/–)	1×10^{-18}	nd	
BM325245	AP003791	157.1	37213–38806 (+/–)	1×10^{-55}	Yes	82G24
BG101800	AP003791	157.1	65825–66170 (+/–)	1×10^{-64}	Yes	155a6
BE598591	AP003791	157.1	91944–92509 (+/–)	7×10^{-69}	Yes	19d6
BG817523	AP003380	157.6	2982–4098 (+/–)	8×10^{-50}	Yes	120h5
BE358320	AP003380	157.6	31795–32165 (+/+)	2×10^{-75}	Yes	84P1
AW283346	AP003416	157.6	12537–12880 (+/–)	1×10^{-36}	Yes	84P1
BG946972	AP003416	157.6	76205–76570 (+/–)	1×10^{-30}	Yes	84P1
BG356230	AP003436	157.6	30570–31084 (+/–)	2×10^{-38}	Yes	180g10
BG356592	AP003436	157.6	79599–80737 (+/+)	9×10^{-32}	Yes	88O4
BM322880	AP003436	157.6	122228–123428 (+/+)	2×10^{-56}	Yes	81G20
BM324063	**AP003436**	**157.6**	**169609–170027 (+/+)**	**2×10^{-50}**	**No**	**53g6**
BG465378	AP003433	158.2	34615–34822 (+/+)	4×10^{-67}	Yes	81G20
BE918371	AP003433	158.2	88409–90208 (+/–)	6×10^{-39}	Yes	133c1
BM322255	**AP003433**	**158.2**	**136689–137065 (+/–)**	**3×10^{-37}**	**No**	**151d9**
AW745139	AP003451	158.5	91287–92335 (+/–)	9×10^{-72}	Yes	133c1
BF656904	AP003451	158.5	149370–146894 (+/–)	4×10^{-40}	Yes	181g10
BG159525	AP003437	159.0	83637–84457 (+/–)	1×10^{-130}	Yes	250h7
BM325684	AP003437	159.0	104319–105317 (+/+)	2×10^{-13}	Yes	250h7
BE361743	AP003707	159.6	49661–50235 (+/–)	2×10^{-47}	Yes	96L18
BG464339	AP004672	159.6–161.5	30718–31785 (+/+)	8×10^{-29}	Yes	88e11
BG649801	AP004672	159.6–161.5	128942–129529 (+/–)	1×10^{-61}	Yes	171e5

[a]Accession number of rice chromosome 1 BAC/PAC with sequence homology to the sorghum EST as determined by BLASTN analysis to the non-redundant database.
[b]Indicates the start and end positions of the region within each rice BAC/PAC sharing homology to the sorghum EST and the nucleotide strands sharing homology.
[c]Denotes whether the sorghum BACs positive for each EST from PCR screening were located on sorghum chromosome 3 and collinear with the corresponding rice BAC/PAC from chromosome 1; nd = not determined because EST primers failed to amplify sorghum genomic DNA; n/a = not available because EST primers amplified a band in all BAC DNA pools.
[d]One of the individual sorghum BACs identified from screening the BAC DNA pools with primers for each EST. Sorghum BACs with a lower case letter in the name (e.g. 111e10) are from the BTx623 libraries and BACs with an upper case letter in the name (e.g. 77G16) are from the IS3620C library.

Twenty-six of the 29 sorghum EST primer sets produced strong PCR signals in the BAC DNA pools that could be easily analyzed, while three primer sets either amplified a band from every DNA pool (ESTs BG103709 and BE598683) or failed to amplify a product even from sorghum genomic DNA (EST AW746672) (Table 3). The 26 sorghum EST primer sets identified many of the sorghum BACs corresponding to the six contigs previously aligned to this region of rice chromosome 1 by sequence scanning (Table 3). However, an additional 31 sorghum BAC clones from the BTx623 libraries as well as 74 clones from the newly constructed IS3620C library were also identified by this PCR-based screening approach.

Ninety-five of the BTx623 and IS3620C BAC clones identified either by PCR-based screening or sequence scanning were subjected to a modified version of high-information content fingerprint (HICF) analysis (Ding *et al.*, 2001; Luo *et al.*, 2003). Following automated contig assembly at an initial cut-off value of 10^{-36}, 84 of the 95 sorghum BAC clones were in two large contigs. By incrementally raising the cut-off to 5×10^{-31}, an additional three singleton clones were added to these two contigs and the two large contigs were merged. The tiling path of the BAC clones spanning sorghum chromosome 3 from approximately 169.2 to 180.7 cM is shown in Figure 2. The alignment between the sorghum and rice physical maps, based on a portion of the results obtained with sequence scanning and PCR screening, is also shown. The contiguous sorghum contig spans approximately 11 cM on sorghum chromosome 3 and encompasses nine markers on the genetic map.

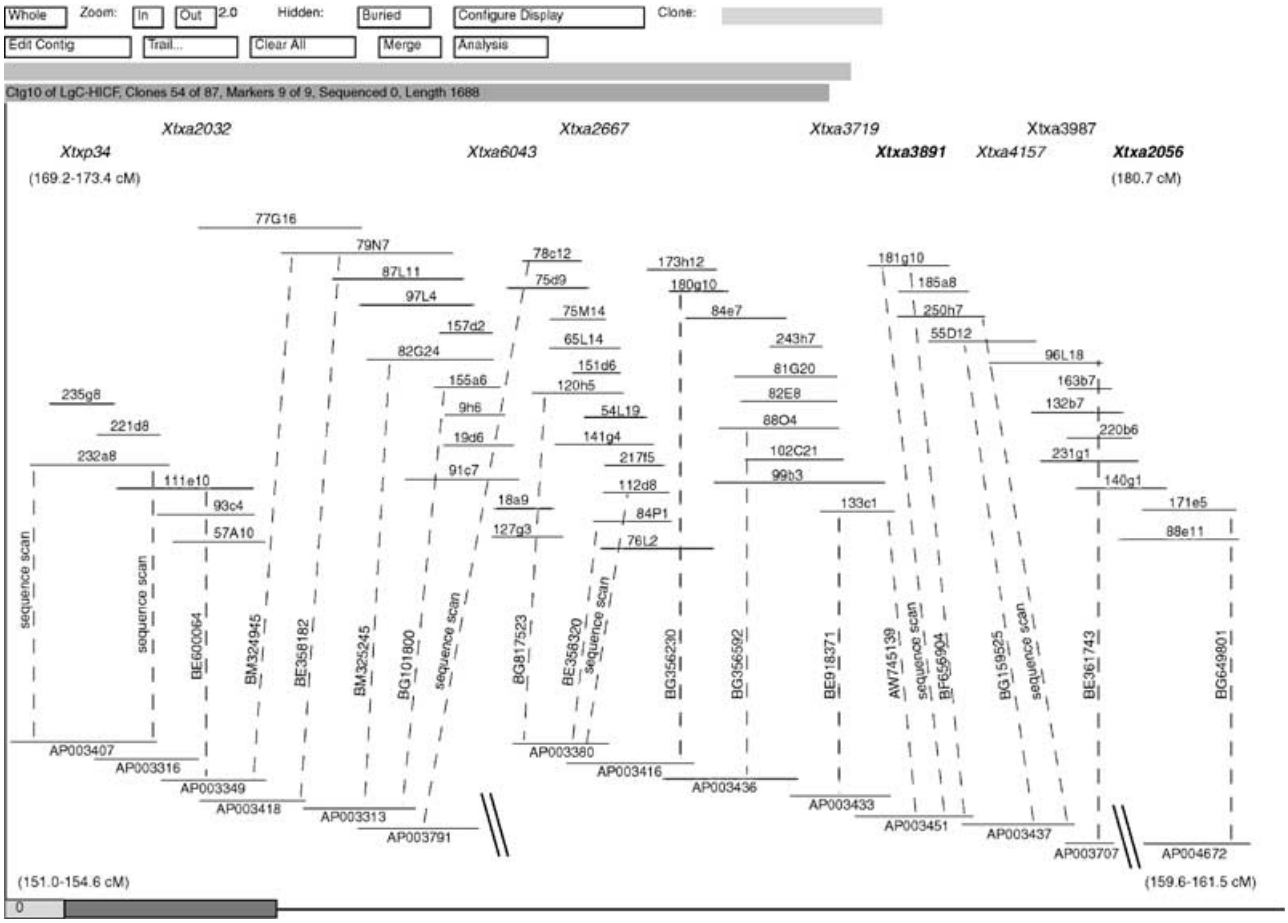

Figure 2. Display of the approximately 1.6 Mbp sorghum contig and alignment to the orthologous region of rice.
Fifty-four of the 87 sorghum clones placed in this contig are shown along with the nine genetic markers encompassing this region of the integrated genetic/physical map of sorghum. Sorghum BACs with a lower case letter in the name (e.g. 111e10) are from the BTx623 libraries, and BACs with an upper case letter in the name (e.g. 77G16) are from the IS3620C library. Framework markers from the high-density sorghum genetic map are shown in bold text, whereas markers 'placed' in bins between framework markers are shown in plain text. The 14 rice chromosome 1 BACs/PACs from the RGP minimum tiling path (http://rgp.dna.affrc.go.jp/) are shown below the sorghum contig. Dashed lines between the sorghum BACs and the rice BACs/PACs denote orthologous regions between the two based on sequence scan data or PCR-based screening with primers to sorghum ESTs. The sorghum ESTs that are orthologous to putative genes in the rice BACs/PACs are indicated by their GenBank accession numbers.

Additionally, the sorghum contig spans the two small gaps present in the rice minimum tiling path in this region, indicating that these sorghum clones can be utilized in a similar screening approach to identify the rice BACs needed to bridge these gaps.

It should be noted that two sorghum ESTs with homology to rice chromosome 1 BACs identified BAC clones that were not in the collinear location on sorghum chromosome 3 (BM324063 and BM322255, Table 3, denoted in bold). As shown in Table 3, the gene corresponding to EST BM324063 lies between the genes corresponding to ESTs BM322880 and BG465378 in the rice tiling path. Primers for both ESTs BM322880 and BG465378 amplify a product from sorghum BAC clone, 81G20, as well as from three other BACs, indicating that these two genes reside on a common sorghum BAC (Table 3). In contrast, primers for EST BM324063 did not amplify a band from these four sorghum BACs but did amplify a fragment from other BACs (data not shown). A similar situation was observed with EST BM322255 that lies between ESTs BE918371 and AW745139 (Table 3). These results suggest that some small-scale rearrangements have occurred in this region of sorghum chromosome 3 relative to rice chromosome 1.

Placement of sorghum markers on the high-density genetic map based on alignment of sorghum/rice chromosomes

The construction of a contig spanning sorghum chromosome 3 from approximately 169.2 to 180.7 cM provided an opportunity to resolve the order of seven genetic markers that were binned in this interval based on genetic mapping data (Menz *et al.*, 2002). In a previous work, four markers (*Xtxp34*, *Xtxa6043*, *Xtxa2667*, and *Xtxa2032*) were placed in

the interval between framework markers *Xtxa3448* (approximately 169.2 cM) and *isu91* (approximately 173.4 cM, Menz *et al.*, 2002). BACs containing each of the four markers within this interval had been identified by screening the BTx623-derived BAC DNA pools. Therefore, once the order of BACs in the contig spanning this region was established, the relative order of these markers was clear (Figure 2). Similarly, the order of binned markers *Xtxa3719, Xtxa4157*, and *Xtxa3987* around framework markers *Xtxa3891 (*175.1 cM) and *Xtxa2056* (180.7 cM) was also established (Figure 2).

Additionally, the BAC sequence data allowed us to correctly order other groups of markers that had been placed into bins within the framework map of sorghum chromosome 3 (Menz *et al.*, 2002 and http://SorghumGenome.tamu.edu). Examination of the sequence-based alignment of sorghum BACs to the rice chromosome 1 map allowed us to tentatively order 13 additional markers mapped to seven different bins. These markers are located in bins between positions 43.0–47.2, 76.5–79.9, 84.7–88.6, 125.6–127.9, 140.4–142.9, 145.9–147.7, and 147.9–150.7 cM on the sorghum chromosome 3 genetic map (Table 2).

Discussion

By the end of 2002, the International Rice Genome Sequencing Project (IRGSP) had completed sequencing the 430 Mbp rice genome to at least phase 2 level (high-quality draft, Sasaki *et al.*, 2002; http://rgp.dna.affrc.go.jp/). The rice genome exhibits substantial collinearity with the genomes of other grasses, such as sorghum, maize, wheat, and barley (Ahn *et al.*, 1993; Chen *et al.*, 1997; van Deynze *et al.*, 1995; Gale and Devos, 1998; Tarchini *et al.*, 2000). This suggests that alignment of the rice genome sequence to high-resolution integrated genetic and physical maps of related species will accelerate the isolation and analysis of genes of agronomic importance from these plants.

Many RFLP-based comparative mapping studies performed between members of the Poaceae have shown that the relative order of markers within extended portions of rice and other grass genomes is conserved (reviewed in Devos and Gale, 1997). More recently, BACs containing orthologous genes from several grass species have been identified and sequenced. Although these studies revealed that most genes within the aligned regions of different grass genomes analyzed were present in the same relative order, small-scale rearrangements including gene insertions, deletions, duplications, and inversions were observed (for review, see Bennetzen, 2000; Dubcovsky *et al.*, 2001; Keller and Feuillet, 2000; Song *et al.*, 2002). These results suggest that utilization of the rice genome sequence for gene discovery in other grasses will require maps with higher resolution alignment than the current RFLP-based comparative genetic maps, and that improved strategies for gene discovery and isolation from larger, more complex genomes are needed.

High-resolution mapping and identification of orthologous sorghum/rice gene sequences

In the present study, we initiated work on high-resolution alignment of our genetically anchored sorghum physical map to the rice genome sequence. In the process, a streamlined approach for the construction and alignment of integrated genetic and physical maps to reference genomes was developed. An essentially complete sequence of rice chromosome 1 has been obtained and annotated by sequencing 390 BACs/PACs from the physical map of this chromosome (Sasaki *et al.*, 2002; http://rgp.dna.affrc.go.jp/cgi-bin/statusdb/status.pl). Therefore, our initial efforts focused on creating high-resolution sequence-based alignments between rice chromosome 1 and our genetic/physical map of sorghum chromosome 3. To this end, we sequence-scanned 118 sorghum BACs mapped to 109 loci on sorghum chromosome 3.

Efficient collection of sequence information from the mapped BACs at low cost required a sequencing strategy that provided sufficient information for alignment and one that could be implemented as a high-throughput method with limited technical difficulty. BAC DNA shearing followed by shotgun sequencing is the preferred method for obtaining complete BAC sequences and a random distribution of sequences across a given genomic region. However, we found that sequencing a limited number of subclones from restriction-digested BAC DNA was the easiest, economical, and most informative way to obtain sequence information for genome alignment. On average, approximately 34% of the sequences obtained from each BAC clone derived from the euchromatic portion of the sorghum genome showed significant similarity with putative or known genes (Table 1). Therefore, a sequence scan of 16–32 subclones from each sorghum BAC from these regions of chromosome 3 provided a good probability of including sequences from one or more genes useful for alignment to the rice genome sequence. This result is consistent with the analysis of several complete sorghum BAC sequences derived from euchromatic regions of the sorghum genome that encode approximately one gene per 5–10 kbp (Morishige *et al.*, 2002; Song *et al.*, 2002; Tikhonov *et al.*, 1999).

The criteria for identifying putative orthologs for genome alignment among sample sequences from sorghum BACs and the rice genome sequence relied on a combination of BLASTX and BLASTN analysis. It has been suggested that nucleotide or amino acid sequence identity is inadequate for identifying orthologs from highly divergent species (Tatusov *et al.*, 1997). However, in closely related taxa, DNA and amino acid sequence comparisons have been

highly successful in identifying orthologous genes (Kim *et al.*, 2001; Thomas *et al.*, 2000). Even between *Arabidopsis* and rice, which diverged from a common ancestor between 120 and 200 million years ago (Wolfe *et al.*, 1989), sequence-based comparisons at the nucleotide and amino acid level have successfully identified small regions of genome collinearity (Liu *et al.*, 2001; Salse *et al.*, 2002). In the present study, BLASTX analysis was used to identify sorghum BAC sequences having similarity with putative or known protein-coding genes whereas BLASTN analysis was used to identify sorghum BAC sequences having sequence similarity with mapped rice BACs. Only those sorghum BAC sequences that shared a similarity with both protein-coding genes and the sequences of mapped rice BACs were analyzed further and used for alignment.

Alignment of sorghum chromosome 3 and rice chromosome 1

Sixty-three of the 118 BACs mapped to sorghum chromosome 3 shared significant sequence similarity with both protein-coding genes and rice chromosome 1 BAC sequences (Table 2). Unique sequences from 59 of these 63 sorghum BACs identified over 100 genes that are present in rice BACs that map in the same relative order on rice chromosome 1, providing substantial alignment between the two chromosomes. After allowing for one major rearrangement, most of the sorghum–rice ortholog pairs were found in the same relative order on sorghum chromosome 3 and rice chromosome 1 (Table 2 and Figure 1). An inversion at the proximal end of sorghum chromosome 3 from approximately 0 to 58.5 cM was observed relative to rice. The location of the breakpoint in sorghum chromosome 3 could be narrowed to an approximately 3.5 cM region between positions 58.5 and 62.0 cM; however, complete sequencing of BACs from this region will be required to delineate the exact position of the breakpoint. Chromosomal rearrangements involving entire arms or segments of arms are quite common among the grasses (Hulbert *et al.*, 1990; Moore *et al.*, 1995; Paterson *et al.*, 1996). In fact, the inversion observed here involving a segment of the short arm of sorghum chromosome 3 was previously reported by Paterson *et al.* (1995) and Ventelon *et al.* (2001). While the large inversion on the short arm of sorghum chromosome 3 was easily detected from our sequence scan, small-scale rearrangements might not have been detected as the average spacing of mapped BACs subjected to sequence scanning was approximately 2–5 cM.

A limited number of studies have compared the complete sequences of aligned genomic regions from members of the Poaceae in order to examine localized conservation of gene content and order (for review, see Bennetzen, 2000). These studies revealed that, in general, a large number of the genes in the regions analyzed were collinear (Avramova *et al.*, 1996; Chen *et al.*, 1997; Dubcovsky *et al.*, 2001; Feuillet *et al.*, 2001; Tikhonov *et al.*, 1999). However, exceptions to collinearity were documented as well (Bancroft, 2001; Dubcovsky *et al.*, 2001; Song *et al.*, 2002; Tarchini *et al.*, 2000; Tikhonov *et al.*, 1999). In a recent study by Song *et al.* (2002), a large orthologous region surrounding the *php20075* marker was characterized in maize, sorghum, and rice. Numerous exceptions to gene collinearity within these orthologous regions were observed, with the majority of the changes resulting from gene amplification, gene movement, and retrotransposition (Song *et al.*, 2002). These authors speculate that gene movement is associated with gene amplification, a phenomenon that has been documented by others as well (Song *et al.*, 2001; Tarchini *et al.*, 2000; Tikhonov *et al.*, 1999). The results of our low-pass sequence scan of sorghum BACs provided evidence for the movement of 15 sequences within sorghum chromosome 3 as well as the movement of 12 sequences to or from sorghum chromosome 3 relative to rice chromosome 1 (Tables 1 and 2). This evidence for intra- and interchromosomal movement of genes is in agreement with the studies of Song *et al.* (2002) and Tikhonov *et al.* (1999). Our results suggest that, in many cases, a single gene may have moved from one location to another within the sorghum genome. However, we also observed movement and/or duplication of larger chromosomal segments containing several genes. For example, sorghum BAC 122c5 exhibited sequence similarity with three genes from the collinear rice BACs AP003339 and AP003282 that are located at 16.4 cM on the rice chromosome 1 genetic map. In addition, sequences from this BAC shared similarity with three genes from a non-collinear rice BAC, AP003768, located at 127.3 cM on the rice chromosome 1 map as well as sequence similarity with a gene on rice BAC, AP002899, located at 126.2 cM on the rice map (Tables 1 and 2). These results may indicate that several rearrangements have occurred in this region of the sorghum genome relative to rice.

Using the rice genome sequence to aid gap filling in the sorghum physical map

If rice chromosome 1 and sorghum chromosome 3 have similar coding capacity and gene density in their euchromatic regions, then gaps in the sorghum physical map will often contain genes that can be identified based on the rice genome sequence. This situation, combined with high-resolution alignment of sorghum chromosome 3 and rice chromosome 1, suggested that a targeted approach to physical map gap filling could be developed. To test this idea, information from the rice genome sequence was used to help in filling six gaps in a targeted region of the sorghum physical map (position approximately 169.2–180.7 cM), resulting in the rapid construction of an approximately

1.6 Mbp sorghum contig. Based on recombination distances from our high-resolution sorghum genetic map, this sorghum BAC contig spans a distance of approximately 11 cM and encompasses nine genetic loci, which were precisely ordered on the physical map (Menz *et al.*, 2002). Our approach was to align sorghum BACs located in the approximately 11 cM region of sorghum chromosome 3 to the rice chromosome 1 physical map by sequence scanning (Tables 1 and 2) and then to identify rice genes and the corresponding orthologous sorghum ESTs located in gaps of the sorghum physical map. PCR primers to these sorghum EST sequences were then used to screen BAC DNA pools from two different sorghum libraries (Table 3). The BAC clones identified by sequence scanning and this PCR-based screening approach were subjected to a newly devised method of HICF analysis for contig construction (Ding *et al.*, 2001; Luo *et al.*, 2003). In this modified HICF method, BAC DNA is simultaneously digested with four different 6-base (*Eco*RI, *Bam*HI, *Xho*I, and *Xba*I) and one 4-base (*Hae*III) recognition restriction enzymes. Each of the four 6-base recognition enzymes produces a different 1-base 5′ overhang that is subsequently filled in using one of the four fluorescent dideoxy terminators in the SNaPshot™ labeling kit. Each product is therefore characterized by both the restriction enzyme producing the fragment and the size of the fragment. The HICF method produces four times more information than standard single restriction enzyme fingerprinting, resulting in a significant increase in the accuracy and efficiency of detecting BAC clones with shared fragments (Ding *et al.*, 2001). It has been estimated that HICF requires less than 20% overlap to construct contigs and, as such, requires 10-fold fewer clones to be fingerprinted than a system requiring 80% overlap to achieve the same level of closure (Ding *et al.*, 2001).

The sequence-based, homology-driven approach used in the present study was successful in rapidly identifying sorghum BAC clones that spanned the targeted 11 cM region of sorghum chromosome 3. HICF analysis of these BAC clones placed them into one contiguous array of overlapping clones (Figure 2). More than 91% of the BAC clones that were identified (87/95) by this PCR-based screening approach were ordered within the approximately 1.6 Mbp contig, indicating a false-positive rate for BAC pool screening of less than 9%. Twenty-four genes were identified within this region and appeared collinear between sorghum and rice. In addition, two sorghum ESTs that were identified by sequence similarity with rice chromosome 1 BACs were not located in the collinear sorghum chromosome 3 BACs, indicating that a small rearrangement had occurred in this region since the divergence of sorghum and rice (Table 3). At present, we do not know the genetic location of these two sorghum genes because the BACs encoding them have not been linked to our genetic map (data not shown). Finally, the sorghum contig spanned two small gaps present in the rice minimum tiling path of BAC clones (Sasaki *et al.*, 2002). The sorghum clones spanning these gaps could be utilized in a similar screening approach to identify the rice BACs needed to bridge these gaps.

The strategy utilized here to construct an extended BAC contig using the rice genomic sequence is similar to the approach being utilized for the construction of a sequence-ready physical map of the mouse genome (Kim *et al.*, 2001; Thomas *et al.*, 2000). In those studies, the human genome sequence was used as a reference to identify orthologous mouse ESTs and a hybridization-based approach based on overgo probes was used to identify the corresponding orthologous mouse BACs (Kim *et al.*, 2001; Thomas *et al.*, 2000). In our study, a PCR-based approach was used to screen six-dimensional pools of BAC DNA to identify collinear sorghum BACs. This approach is efficient, cost-effective, requires minimal technical skill, and can easily be implemented at high throughput. This screening approach has been used in combination with AFLP technology to successfully link more than 1100 sorghum genetic markers to BAC clones (Klein *et al.*, 2000; P.E. Klein, unpublished data). Additionally, with the construction of the new BAC library derived from IS3620C, the second parent of our RIL mapping population, this methodology can be used to generate another 1000 links to the high-density sorghum genetic map (Menz *et al.*, 2002 and http://SorghumGenome.tamu.edu).

Gene density and the overall architecture of sorghum chromosome 3

It has been hypothesized that, in rice, the majority of genes are found in regions of relatively high gene density accounting for approximately 12–24% of the genome separated by large gene-poor regions (Barakat *et al.*, 1997). The results of sequence scanning sorghum BACs mapped at intervals across chromosome 3 suggest that gene density is higher in the distal regions of this chromosome compared to the pericentromeric region centered at approximately 76.5–79.9 cM (Table 2 and Figure 1). These results are consistent with those obtained by Gomez *et al.* (1997), Zwick *et al.* (2000), and Islam-Faridi *et al.* (2002), suggesting that gene density is highest in the distal euchromatic regions of sorghum chromosomes. Similar results have also been reported for wheat and barley (Gill *et al.*, 1996a,b; Kuenzel *et al.*, 2000). The sorghum BACs that mapped to the pericentromeric region of chromosome 3 were largely devoid of gene sequences but enriched in repetitive elements and genes associated with transposable elements. Additionally, this region exhibits suppressed genetic recombination with approximately 83 loci mapping to an approximately 4 cM region around the sorghum chromosome 3 centromere

(Menz *et al.*, 2002). It is likely that this region of sorghum chromosome 3 encompasses a large segment of DNA, as observed with other sorghum chromosomes (Islam-Faridi *et al.*, 2002). Similarly, in rice, it has been shown that the centromeres are located in extended regions of heterochromatin that exhibit suppressed genetic recombination and encompass large physical distances (Chen *et al.*, 2002). These results indicate that physical mapping and alignment of the pericentromeric regions of sorghum chromosomes will be difficult; however, if these regions are largely devoid of genes, most of the initial gene discovery projects will occur in euchromatic regions.

Conclusions

In summary, we have developed an efficient strategy for alignment of the sorghum and rice genome maps. Our results indicate that the overall architecture of sorghum chromosome 3 and rice chromosome 1 has remained largely intact with the exception of one major rearrangement. This is consistent with early comparative mapping studies based on low-copy RFLP probes (Paterson *et al.*, 1995; Peng *et al.*, 1999; Ventelon *et al.*, 2001). However, there has also been a significant amount of movement or reorganization of single genes or groups of genes within sorghum chromosome 3 and between sorghum chromosome 3 and other parts of the sorghum genome relative to rice. This observation is consistent with analysis of the complete sequences of collinear BACs from different grass genomes that show localized gene movements, inversions, deletions, and other changes (for review, see Bennetzen, 2000; Dubcovsky *et al.*, 2001; Keller and Feuillet, 2000; Song *et al.*, 2002). This situation allows the overall conservation of gene order along sorghum chromosome 3 and rice chromosome 1 to be used for construction and alignment of collinear maps, but also emphasizes the need for relatively deep sequence analysis of sorghum if the value of comparative genomics is to be fully realized. It is our intent that methods developed in this study will yield important information about the relationships between rice and sorghum genomic segments, provide a well-characterized clone set for sequencing the euchromatic portion of the sorghum genome, and ultimately enhance our understanding of cereal genome structure and evolution.

Experimental procedures

Mapping genetic markers to BAC DNA

AFLP mapping. As detailed elsewhere (Klein *et al.*, 2000), DNA pools were constructed from two sorghum BAC libraries made from the elite sorghum genotype, BTx623. AFLP template, *Eco*RI/*Mse*I and *Pst*I/*Mse*I, was prepared from the resulting 184 pools as described (Klein *et al.*, 2000; Menz *et al.*, 2002). AFLP pre-amplification (0/+1 primers) and selective amplification (+3/+3 primers) reactions of BAC pools were performed as described (Klein *et al.*, 2000; Menz *et al.*, 2002), using the 192 AFLP primer combinations previously used for the development of the high-density AFLP-based genetic map of sorghum (Menz *et al.*, 2002). Amplification products were analyzed as described (Klein *et al.*, 2000). Individual BAC clones harboring AFLP genetic markers were identified from the BAC DNA pools using a Unix-based application written in the Perl programming language (Klein *et al.*, 2000). When analyzing candidate BAC clones for the presence of a marker, the selective amplification reaction was modified as described (Klein *et al.*, 2000).

SSR and RFLP mapping. BTx623 BAC DNA pools were screened for sorghum SSRs or RFLPs previously mapped to chromosome 3 as described (Klein *et al.*, 2001). Individual clones containing an SSR or RFLP were identified from the BAC DNA pools using the Unix-based application as described above.

Sample sequencing of BAC DNA

One hundred and eighteen individual BAC clones from the BTx623 library were chosen for sample sequencing based on their linkage to sorghum chromosome 3 genetic markers (AFLPs, SSRs, and/or RFLPs). BAC DNA was prepared from a 5 ml culture (LB medium containing 12.5 μg ml^{-1} chloramphenicol) by alkaline lysis purification. Following isopropanol precipitation, BAC DNA was re-suspended in 40 μl of sterile water and mixed with 10 μl of cocktail containing 10 U of Plasmid-safe ATP-dependent DNase (Epicentre Technologies, Madison, WI, USA), 1 μl of RNase cocktail (Ambion, Austin, TX, USA), 1 mM ATP, and 5 μl of 10$\times$ Plasmid-safe buffer supplied by the manufacturer. The reactions were incubated overnight at 37°C followed by DNase inactivation by incubation at 70°C (30 min). BAC DNA was digested with *Eco*RI and *Xho*I by adding 50 μl of enzyme cocktail containing 10 U of *Eco*RI (New England Biolabs, Beverly, MA, USA), 10 U of *Xho*I (Invitrogen, Carlsbad, CA, USA), 50 mM Tris–HCl (pH 8.0), 10 mM $MgCl_2$, and 50 mM NaCl. The reactions were incubated at 37°C for 3–4 h and the fragments purified using the Qiagen QIAquick PCR purification kit according to the manufacturer's protocol (Qiagen, Valencia, CA, USA). Purified, digested BAC DNA (4 μl) was ligated into *Eco*RI/*Xho*I-digested pBluescript SK$^+$ (Stratagene, La Jolla, CA, USA) for 16 h at 16°C. Ligation reactions were transformed into chemically competent DH5-α subcloning efficiency cells and plated on LB agar plates containing 100 μg ml^{-1} ampicillin.

For each BAC, 16–96 subclones were picked into deep-well microtiter plates containing 1.4 ml of LB media containing 100 μg ml^{-1} ampicillin and grown overnight (37°C, 325 r.p.m.). Plasmid DNA was purified on a Biomek 2000 (Beckman Coulter, Fullerton, CA, USA) using a Wizard MagneSil plasmid purification kit according to the manufacturer's protocol (Promega, Madison, WI, USA). Sequencing was performed with either the reverse (5′-GGAAACAGCTATGACCATG-3′) or T3 (5′-AATTAACCCTCACTAAAGGG-3′) primers. Standard sequencing reactions included 2 μl of plasmid DNA, 10 pmol of primer, 0.66 μl of BigDye Terminator mix v2.0 (Applied Biosystems, Foster City, CA, USA), 1.33 μl of 5$\times$ reaction buffer (5$\times$ = 400 mM Tris, pH 9.0, 10 mM $MgCl_2$), and 5 μl of distilled water. Sequencing reactions were carried out using the following cycling parameters: an initial denaturation at 95°C for 2 min, followed by 99 cycles of 95°C for 10 sec, 50°C for 5 sec, and 60°C for 4 min. Extension products were purified by isopropanol precipitation and separated on an ABI3700 DNA sequencer. Sequence files were processed using PHREDPHRAP,

and sequence homology searches were performed against the non-redundant nucleotide and protein databases (BLASTN and BLASTX, respectively) and the htgs database (BLASTN, Altschul *et al.*, 1997) using a local BLAST server. Following BLAST analysis, the data was parsed to eliminate those sequences with limited homology to the BAC subclones (*E*-value $< 10^{-5}$) and the results were imported into a MySQL database for further analysis.

Preparation of high-molecular-weight Sorghum bicolor *DNA and BAC library construction*

A third sorghum BAC library was constructed for this study using the converted sorghum line IS3620C. This converted line is one of the parents of the recombinant inbred mapping population used to construct the sorghum high-density genetic map (BTx623 × IS3620C, Menz *et al.*, 2002). Extraction of high-molecular-weight genomic DNA from leaf nuclei was performed as described in Zhang *et al.* (1995) and partially digested with *Hin*dIII followed by three rounds of size selection after separation by pulsed field gel electrophoresis (PFGE). Prior to ligation, the final size-selected DNA (>150 kb) was released from agarose by electroelution for 2 h at 200 V with a 90 sec pulse at 11°C. Eluted DNA was quantified on a gel, and a molar ratio of approximately 3 : 1 vector:insert was used for ligation. pBelobacII (Shizuya *et al.*, 1992) that had been *Hin*dIII-digested and dephosphorylated was used for library construction. Transformations were performed by electroporation using GeneHogs electrocompetent cells (Invitrogen, Carlsbad, CA, USA). Approximately 44 000 clones were selected following transformation and plating. The average insert size was estimated at 170 kb following *Not*I digestion and gel electrophoresis.

Six-dimensional BAC pooling and mapping sorghum ESTs to BAC DNA

A subset of 24 576 BAC clones from the IS3620C *Hin*dIII library was pooled on six coordinate axes and BAC DNA isolated from the resulting 184 pools as described (Klein *et al.*, 2000).

Primers to sorghum ESTs were designed using the OLIGO 6.0 software program (Molecular Biology Insights Inc., Cascade, CA, USA) and were obtained from Sigma–Genosys (The Woodlands, TX, USA). PCR-based screening of BTx623 and IS3620C BAC DNA pools was performed in 10 μl reactions containing 1× Perkin-Elmer buffer II (Applied Biosystems, Foster City, CA, USA), 2.5 mM $MgCl_2$, 200 μM dNTPs, 0.4 U of AmpliTaq polymerase, 20 pmol of each primer, and 5 ng of pooled BAC DNA or genomic DNA (BTx623 and IS3620C). Amplification conditions were as previously described (Klein *et al.*, 1998). Amplification products were electrophoresed in 2% agarose gels and the products visualized following staining with SYBR Gold (Molecular Probes Inc., Eugene, OR, USA). BAC DNA pools containing signals for a given sorghum EST were recorded, and the individual BACs containing the EST were identified using the Unix Perl script as described above.

DNA fingerprinting of BAC clones

BAC clones were fingerprinted using a modified version of five-color-based high-information content fingerprinting (Luo *et al.*, 2003). BAC DNA was isolated in deep-well plate format as previously described, except for the addition of RNaseA (100 μg ml^{-1}) to the initial re-suspension buffer (Klein *et al.*, 1998). Following purification, BAC DNA pellets were re-suspended in 45 μl sterile water and quantified by fluorimetry, and the DNA concentration was adjusted to 50–75 ng ml^{-1} with sterile water. BAC DNA was restricted by mixing 42 μl of purified DNA with 9 μl of enzyme cocktail consisting of 5 μl of 10× NEBuffer 2 (New England Biolabs, Beverly, MA, USA), 0.5 μl of BSA (100 mg ml^{-1}), 1 μl of RNaseA (0.5 μg ml^{-1}), 1 μl of 1% α-mercaptoethanol, 5 U of *Bam*HI, 5 U of *Eco*RI, 5 U of *Xba*I, 5 U of *Xho*I, and 5 U of *Hae*III. The reactions were incubated for 3 h at 37°C. Following restriction enzyme digestion, 10 μl of labeling cocktail containing 2 μl of 10× NEBuffer 2, 2.5 μl of Tris, pH 9.0, 1 μl SNaPshot™ Multiplex Ready Reaction Mix (Applied Biosystems), and 4.5 μl of sterile water was added to each sample, and the reactions were incubated at 65°C for 1 h. Reactions were precipitated by the addition of 5.0 μl of 2.5 M sodium acetate and 100 μl of chilled 95% ethanol (− 80°C, 15 min). Labeled fragments were collected by centrifugation at 1920 ***g*** for 30 min, washed once with 70% ethanol, and air-dried. Pellets were re-suspended in 9.75 μl of Hi-Di formamide containing 0.25 μl of GeneScan™-500 Liz size standard (Applied Biosystems). Labeled fragments were separated on an ABI3700 sequencer using the SNP2_POP5 module with modifications. Electrophoresis was performed at 6000 V at a run temperature of 50°C for 90 min. Collected data was analyzed with GeneScan™ version 3.7 Fragment Analysis Software (Applied Biosystems). The minimum peak detection threshold was individually set for each lane and for each fluorescent dye. Data was manually edited using Genotyper™ version 3.7 Fragment Analysis Software (Applied Biosystems), and a table consisting of a list of fragment sizes and dye colors for each BAC clone was exported in a tabular format. Conversion of the Genotyper™ output table into a band file for input into FPC was accomplished using a script written in the Perl programming language. Automated contig assembly was performed using FPC V4.9 (Soderlund *et al.*, 1997) at a fixed tolerance of one and an initial cut-off value of 10^{-36}. The cut-off was subsequently raised to 5×10^{-31} for the addition of singletons to existing contigs and the merging of contigs.

Acknowledgements

The authors wish to thank Dr M.C. Luo and Dr Jan Dvora (University of California, Davis) for graciously providing the protocol for the modified HICF fingerprinting method prior to its publication and for answering all our technical questions regarding this protocol. The authors also thank Julie McCollum and Jacque Obert for technical assistance and Soumaya Fassi-Fehri for development of the MySQL database. This work was supported in part by an NSF Plant Genome Research Grant DBI-0077713 (J.E.M. and P.E.K.) and the USDA's Agricultural Research Service (R.R.K.).

References

Adams, M.D., Celniker, S.E., Holt, R.A. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.

Ahn, S., Anderson, J.A., Sorrells, M.E. and Tanksley, S.D. (1993) Homoeologous relationships of rice, wheat and maize chromosomes. *Mol. Gen. Genet.* **241**, 483–490.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.

Arumuganathan, K. and Earle, E.D. (1991) Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.* **9**, 208–219.

Avramova, Z., Tikhonov, A., SanMiguel, P., Jin, Y.-K., Liu, C.-N., Woo, S.-S., Wing, R.A. and Bennetzen, J.L. (1996) Gene identification in a complex chromosomal continuum by local genomic cross-referencing. *Plant J.* **10**, 1163–1168.

Bancroft, I. (2001) Duplicate and diverge: the evolution of plant genome microstructure. *Trends Genet.* **17**, 89–93.

Barakat, A., Carels, N. and Bernardi, G. (1997) The distribution of genes in the genomes of Gramineae. *Proc. Natl. Acad. Sci. USA*, **94**, 6857–6861.

Bennetzen, J.L. (2000) Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. *Plant Cell*, **12**, 1021–1029.

Binelli, G., Gianfranceschi, L., Pe, M.E., Taramino, G., Busso, C., Stenhouse, J. and Ottaviano, E. (1992) Similarity of maize and sorghum genomes as revealed by maize RFLP probes. *Theor. Appl. Genet.* **84**, 10–16.

Chen, M., Presting, G., Barbazuk, W.B. *et al.* (2002) An integrated physical and genetic map of the rice genome. *Plant Cell*, **14**, 537–545.

Chen, M., SanMiguel, P., de Oliveira, A.C., Woo, S.S., Zhang, H., Wing, R.A. and Bennetzen, J.L. (1997) Microcolinearity in *sh2*-homologous regions of the maize, rice, and sorghum genomes. *Proc. Natl. Acad. Sci. USA*, **94**, 3431–3435.

Cheng, Z., Buell, C.R., Wing, R.A., Gu, M. and Jiang, J. (2001) Toward a cytological characterization of the rice genome. *Genome Res.* **11**, 2133–2141.

Devos, K.M. and Gale, M.D. (1997) Comparative genetics in the grasses. *Plant Mol. Biol.* **35**, 3–15.

van Deynze, A.E., Nelson, J.C., Yglesias, E.S., Harrington, S.E., Braga, D.P., McCouch, S.R. and Sorrells, M.E. (1995) Comparative mapping in grasses. Wheat relationships. *Mol. Gen. Genet.* **248**, 744–754.

Ding, Y., Johnson, M.D., Chen, W.Q., Wong, D., Chen, Y.-J., Benson, S.C., Lam, J.Y., Kim, Y.-M. and Shizuya, H. (2001) Five-color-based high-information-content fingerprinting of bacterial artificial chromosome clones using Type IIS restriction endonucleases. *Genomics*, **74**, 142–154.

Doggett, H. (1988) *Sorghum*, 2nd edn. New York: John Wiley.

Dubcovsky, J., Ramakrishna, W., SanMiguel, P.J., Busso, C.S., Yan, L.L., Shiloff, B.A. and Bennetzen, J.L. (2001) Comparative sequence analysis of colinear barley and rice bacterial artificial chromosomes. *Plant Physiol.* **125**, 1342–1353.

Feuillet, C., Penger, A., Gellner, K., Mast, A. and Keller, B. (2001) Molecular evolution of receptor-like kinase genes in hexaploid wheat. Independent evolution of orthologs after polyploidization and mechanisms of local rearrangements at paralogous loci. *Plant Physiol.* **125**, 1304–1313.

Gale, M.D. and Devos, K.M. (1998) Comparative genetics in the grasses. *Proc. Natl. Acad. Sci. USA*, **95**, 1971–1974.

Gill, K.S., Gill, B.S., Endo, T.R. and Boyko, E.V. (1996a) Identification and high-density mapping of gene-rich regions in chromosome group 5 of wheat. *Genetics*, **143**, 1001–1012.

Gill, K.S., Gill, B.S., Endo, T.R. and Taylor, T. (1996b) Identification and high-density mapping of gene-rich regions in chromosome group 1 of wheat. *Genetics*, **144**, 1883–1891.

Gomez, M.I., Islam-Faridi, M.N., Woo, S.-S., Schertz, K.F., Czeschin, D., Jr, Zwick, M.S., Wing, R.A., Stelly, D.M. and Price, H.J. (1997) FISH of a maize *sh2*-selected sorghum BAC to chromosomes of *Sorghum bicolor*. *Genome*, **40**, 475–478.

Helentjaris, T., Weber, D.L. and Wright, S. (1988) Identification of the genomic locations of duplicate nucleotide sequences in maize by analysis of restriction fragment length polymorphisms. *Genetics*, **118**, 353–363.

Hulbert, S.H., Richter, T.E., Axtell, J.D. and Bennetzen, J.L. (1990) Genetic mapping and characterization of sorghum and related crops by means of maize DNA probes. *Proc. Natl. Acad. Sci. USA*, **87**, 4251–4255.

Islam-Faridi, M.N., Childs, K.L., Klein, P.E., Hodnett, G., Menz, M.A., Klein, R.R., Rooney, W.L., Mullet, J.E., Stelly, D.M. and Price, H.J. (2002) A molecular cytogenetic map of sorghum chromosome *1*: fluorescence *in situ* hybridization analysis with mapped bacterial artificial chromosomes. *Genetics*, **161**, 345–353.

Keller, B. and Feuillet, C. (2000) Colinearity and gene density in grass genomes. *Trends Plant Sci.* **5**, 246–251.

Kim, J., Gordon, L., Dehal, P. *et al.* (2001) Homology-driven assembly of a sequence-ready mouse BAC contig map spanning regions related to the 46-Mb gene-rich euchromatic segments of human chromosome 19. *Genomics*, **74**, 129–141.

Klein, R.R., Morishige, D.T., Klein, P.E., Dong, J. and Mullet, J.E. (1998) High throughput BAC DNA isolation for physical map construction of sorghum (*Sorghum bicolor*). *Plant Mol. Biol. Rep.* **16**, 351–364.

Klein, P.E., Klein, R.R., Cartinhour, S.W. *et al.* (2000) A high-throughput AFLP-based method for constructing integrated genetic and physical maps: progress toward a sorghum genome map. *Genome Res.* **10**, 789–807.

Klein, R.R., Rodriguez-Herrera, R., Schlueter, J.A., Klein, P.E., Yu, Z.H. and Rooney, W.L. (2001) Identification of genomic regions that affect grain-mould incidence and other traits of agronomic importance in sorghum. *Theor. Appl. Genet.* **102**, 307–319.

Kuenzel, G., Korzun, L. and Meister, A. (2000) Cytologically integrated physical restriction fragment length polymorphism maps for the barley genome based on translocation breakpoints. *Genetics*, **154**, 397–412.

Liu, H., Sachidanandam, R. and Stein, L. (2001) Comparative genomics between rice and *Arabidopsis* shows scant collinearity in gene order. *Genome Res.* **11**, 2020–2026.

Luo, M.C., Thomas, C., You, F.M., Hsiao, J., Ouyang, S., Buell, C.R., Malandro, M., McGuire, P.E., Anderson, O.D. and Dvorak, J. (2003) High-throughput fingerprinting of bacterial artificial chromosomes using the SNaPshot™ labeling kit and sizing of restriction fragments by capillary electrophoresis. *Genomics* (in press).

Menz, M.A., Klein, R.R., Mullet, J.E., Obert, J.A., Unruh, N.C. and Klein, P.E. (2002) A high-density genetic map of *Sorghum bicolor* (L.) Moench based on 2926 AFLP®, RFLP and SSR markers. *Plant Mol. Biol.* **48**, 483–499.

Moore, G., Devos, K.M., Wang, Z. and Gale, M.D. (1995) Grasses, line up and form a circle. *Curr. Biol.* **5**, 737–739.

Morishige, D.T., Childs, K.L., Moore, L.D. and Mullet, J.E. (2002) Targeted analysis of orthologous *phytochrome A* regions of the sorghum, maize, and rice genomes using gene-island sequencing. *Plant Physiol.* **130**, 1614–1625.

Panstruga, R., Buschges, R., Piffanelli, P. and Schulze-Lefert, P. (1998) A contiguous 60 kb genomic stretch from barley reveals molecular evidence for gene islands in a monocot genome. *Nucl. Acids Res.* **26**, 1056–1062.

Paterson, A.H., Lin, Y.-R., Li, Z., Schertz, K.F., Doebley, J.F., Pinson, S.R.M., Liu, S.-C., Stansel, J.W. and Irvine, J.E. (1995) Convergent domestication of cereal crops by independent mutations at corresponding genetic loci. *Science*, **269**, 1714–1718.

Paterson, A.H., Lan, T.H., Reischmann, K.P. *et al.* (1996) Toward a unified genetic map of higher plants, transcending the monocot-dicot divergence. *Nat. Genet.* **14**, 380–382.

Peng, Y., Schertz, K.F., Cartinhour, S. and Hart, G.E. (1999) Comparative genome mapping of *Sorghum bicolor* (L.) Moench using an RFLP map constructed in a population of recombinant inbred lines. *Plant Breed.* **118**, 225–235.

Salse, J., Piegu, B., Cooke, R. and Delseny, M. (2002) Synteny between *Arabidopsis thaliana* and rice at the genome level: a

tool to identify conservation in the ongoing rice genome sequencing project. *Nucl. Acids Res.* **30**, 2316–2328.

SanMiguel, P. and Bennetzen, J.L. (1999) Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. *Ann. Bot.* **82**, 37–44.

Sasaki, T., Matsumoto, T., Yamamoto, K. *et al.* (2002) The genome sequence and structure of rice chromosome 1. *Nature*, **420**, 312–316.

Shizuya, H., Birren, B., Kim, U.J., Mancino, V., Slepak, T., Tachiiri, Y. and Simon, M. (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc. Natl. Acad. Sci. USA*, **89**, 8794–8797.

Soderlund, C., Longden, I. and Mott, R. (1997) FPC: a system for building contigs from restriction fingerprinted clones. *Comput. Applic. Biosci.* **13**, 523–535.

Song, R., Llaca, V., Linton, E. and Messing, J. (2001) Sequence, regulation, and evolution of the maize 22-kD á zein gene family. *Genome Res.* **11**, 1817–1825.

Song, R., Llaca, V. and Messing, J. (2002) Mosaic organization of orthologous sequences in grass genomes. *Genome Res.* **12**, 1549–1555.

Tarchini, R., Biddle, P., Wineland, R., Tingey, S. and Rafalski, A. (2000) The complete sequence of 340 kb of DNA around the rice *Adh1-Adh2* region reveals interrupted colinearity with maize chromosome 4. *Plant Cell*, **12**, 381–391.

Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

The *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, **282** (11), 1998.

Thomas, J.W., Summers, T.J., Lee-Lin, S.-Q., Maduro Braden, V.V., Idol, J.R., Mastrian, S.D., Ryan, J.F., Jamison, D.C. and Green, E.D. (2000) Comparative genome mapping in the sequence-based era: early experience with human chromosome 7. *Genome Res.* **10**, 624–633.

Tikhonov, A.P., SanMiguel, P.J., Nakajima, Y., Gorenstein, N.M., Bennetzen, J.L. and Avramova, Z. (1999) Colinearity and its exceptions in orthologous *adh* regions of maize and sorghum. *Proc. Natl. Acad. Sci. USA*, **96**, 7409–7414.

Ventelon, M., Deu, M., Garsmeur, O., Doligez, A., Ghesquiere, A., Lorieux, M., Rami, J.F., Glaszmann, J.C. and Grivet, L. (2001) A direct comparison between the genetic maps of sorghum and rice. *Theor. Appl. Genet.* **102**, 379–386.

Venter, J.C., Adams, M.D., Myers, E.W. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1349.

Wolfe, K.H., Gouy, M., Yang, Y.W., Sharp, P.M. and Li, W.H. (1989) Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. *Proc. Natl. Acad. Sci. USA*, **86**, 6201–6205.

Zhang, H.-B., Zhao, X., Ding, X., Paterson, A.H. and Wing, R.A. (1995) Preparation of megabase-size DNA from plant nuclei. *Plant J.* **7**, 175–184.

Zwick, M.S., Islam-Faridi, M.N., Zhang, H.B., Hodnett, G.L., Gomez, M.I., Kim, J.S., Price, H.J. and Stelly, D.M. (2000) Distribution and sequence analysis of the centromere-associated repetitive element CEN38 of *Sorghum bicolor* (Poaceae). *Am. J. Bot.* **87**, 1757–1764.

Accession numbers: BZ412839–BZ413010.